

Heterogeneous Reciprocal Graphical Models

Yang Ni¹, Peter Müller², Yitan Zhu³, and Yuan Ji^{3,4}

¹Department of Statistics and Data Sciences, The University of Texas at Austin

²Department of Mathematics, The University of Texas at Austin

³Program for Computational Genomics & Medicine, NorthShore University HealthSystem

⁴Department of Public Health Sciences, The University of Chicago

December 20, 2016

Abstract

We develop novel hierarchical reciprocal graphical models to infer gene networks from heterogeneous data. In the case of data that can be naturally divided into known groups, we propose to connect graphs by introducing a hierarchical prior across group-specific graphs, including a correlation on edge strengths across graphs. Thresholding priors are applied to induce sparsity of the estimated networks. In the case of unknown groups, we cluster subjects into subpopulations and jointly estimate cluster-specific gene networks, again using similar hierarchical priors across clusters. We illustrate the proposed approach by simulation studies and two applications in multiplatform genomic data for multiple cancers.

Keywords: Dirichlet-multinomial allocation; hierarchical model; model-based clustering; multiplatform genomic data; Pitman-Yor process; thresholding prior.

1 Introduction

We develop a heterogeneous reciprocal graphical model (HRGM) to infer gene networks in heterogeneous populations. Traditional graphical model approaches (Wang and West, 2009; Dobra et al., 2012; Green and Thomas, 2013; Wang et al., 2013; Peterson et al., 2015) assume i.i.d. sampling. However, many applications to inference for biomedical data, including the applications in this paper, include highly heterogeneous populations, and understanding and characterizing such heterogeneity is an important inference goal. We therefore propose an approach that admits statistical inference for potentially heterogeneous gene regulatory relationships across *known or unknown* groups/subpopulations. In particular, for the case of known groups, we model the related graphs under a Bayesian hierarchical model framework and allow the information to be shared across different groups. Borrowing of strength is implemented for the graph structure as well as for the edge strengths. For the case of unknown groups, we propose to cluster the subjects into subpopulations with meaningfully different graphs, that is, with group-specific graphs that differ in ways that allow biologically meaningful interpretation. Importantly, we implement clustering based on differences in the networks, in contrast to earlier proposed clustering methods (Dahl, 2006; Quintana, 2006; Lau and Green, 2007; Müller et al., 2012; Lijoi et al., 2014), which are mostly based on cluster-specific mean/location.

Our work is motivated by two cancer genomic applications. In the first application, we construct gene networks for three different cancer types. Recent pan-cancer studies (Hoadley et al., 2014) find both differences and commonalities across cancers despite of different tissue-of-origin. Traditional methods that assume homogeneous population are not suitable in this case. Novel statistical methods accounting for data heterogeneity as well as adaptively borrowing strength across subtypes are needed. Our second application finds motivation in breast cancer which is molecularly heterogeneous. Current classification systems based on three biomarkers are argued to be suboptimal as a means of directing therapeutic decisions for breast cancer patients (Di Leo et al., 2015). Improving the classification system is particularly important and yet challenging. This calls for better ways of clustering breast cancer patients.

Graphical models are commonly used to probabilistically model gene regulations. We

use a less commonly used class of graphical models, namely reciprocal graphical models (RGMs, Koster 1996). RGMs are a flexible class of models that allow for undirected edges, directed edges and directed cycles (ideal for modeling biological feedback loops). RGMs strictly contain Markov random fields (MRFs) and directed acyclic graphs (DAGs) as special cases. However, it is surprisingly underused in biostatistics and bioinformatics literature. We prefer to use the RGM over other graphical models, because the inclusion of directed edges and possible cycles is critical for the two motivating applications.

Although graphical models for homogeneous data have been studied extensively in the literature, only few approaches have been proposed for heterogeneous data. When the groups are known, it is natural to “borrow strength” from different sample groups via Bayesian hierarchical modeling in estimating group-specific graphs. We provide a brief review of recent Bayesian methods and discuss the need and possibility for improvement. Mitra et al. (2016) consider MRFs for $K = 2$ groups and name the first group the reference group and the second group the differential group. They assign a uniform prior to the reference graph and construct a mixture prior for the differential graph. Similarly, Oates et al. (2015) develop a multiple DAG approach for $K > 2$. They penalize the difference between graphs based on structural Hamming distance and utilize integer linear programming to find the posterior mode. Peterson et al. (2015) couple undirected MRF graphs by assigning an MRF prior on the edges. Computation can be challenging when K is moderate because of an analytically intractable normalization constant in the MRF model. One common limitation of these methods is that they only borrow strength in the graph space, leaving the strength of selected edges (e.g. partial correlation) to be modeled/estimated independently. One exception is Yajima et al. (2015). They propose a multiple DAG which correlates the edge strength across groups. However, similarly to Mitra et al. (2016), they only consider $K = 2$ groups and inference depends on the choice of reference group and differential group. Generalization to $K > 2$ is not straightforward. For non-Bayesian methods, Guo et al. (2011) and Danaher et al. (2014) develop penalized likelihood approaches to encourage similarities among groups for MRFs.

There are few approaches for unknown groups. Rodriguez et al. (2011) propose a Dirichlet process mixture of MRFs to simultaneously cluster samples into homogeneous groups

and infer undirected relationships between variables for each group. Similarly, Ickstadt et al. (2011) propose a multinomial-Dirichlet-Poisson mixture of DAGs to study directed relationships within each cluster. Mukherjee and Rodriguez (2015) recently develop a GPU-based stochastic search algorithm to improve the computation in Rodriguez et al. (2011).

In this article, we propose a hierarchical extension of RGMs to heterogeneous RGMs (HRGMs) as a model for both cases, known and unknown groups. When groups are known, the HRGM borrows strength across groups for inference on the group-specific graphs as well as the strength of the included edges. For unknown groups, the HRGM clusters a heterogeneous population into homogeneous subpopulations and allows each cluster to have its own network.

The remainder of the paper is organized as follows. In Section 2, we introduce RGMs for modeling multiplatform genomic data. We provide the probability model and prior specifications for known and unknown groups in Sections 3 and 4, respectively. We present simulation studies in Section 5 and two applications in breast cancer in Section 6. Section 7 provides our closing discussion.

2 Reciprocal graphs

A graph $\mathcal{G} = (V, E)$ consists of a set of vertices $V = \{1, \dots, p\}$ usually representing a set of random variables and a set of edges E connecting these vertices. A reciprocal graph admits both directed edges $i \rightarrow j$ and undirected edges $i - j$. Moreover, it explicitly allows for directed cycles, which is useful for modeling feedback mechanism — a common motif in a gene network. Markov properties (i.e. conditional independence relationships) can be read off from the RGM through moralization. Statistically, RGMs are a strictly larger class of probability models than MRFs and DAGs in terms of conditional independence (Koster, 1996).

RGMs are only identifiable up to Markov equivalence class in which all RGMs encode the same conditional independence relationships. Interventional or time-course data or other external information are needed to learn the structure of RGMs and determine the directionality of the edges. Recently, several approaches (Cai et al., 2013; Zhang and Kim,

2014; Ni et al., 2016a) have proposed to infer directionality with observational data by integrating multiplatform data and fixing some directions of edges via biological knowledge. In this paper, we follow the approach of Ni et al. (2016a) to include external information about known directional edges, by exploiting the central dogma of molecular biology, implying that DNA methylation and DNA copy number can affect mRNA gene expression, but not vice versa. Based on this consideration, we integrate DNA copy number and DNA methylation with mRNA gene expression and fix the direction of edges between DNA measurements and mRNA gene expressions. Theoretical justifications for such framework can be found in Logsdon and Mezey (2010) and Oates et al. (2016).

We briefly review the mapping between RGMs and simultaneous equation models (SEMs). Let $\mathbf{y}_i \in \mathbb{R}^p$ denote gene expressions for p genes and $\mathbf{x}_i \in \mathbb{R}^q$ denote q DNA level measurements (copy number and methylation in our case study) for subject $i = 1, \dots, n$. Consider $(\mathbf{y}_i, \mathbf{x}_i)$ to be jointly normal and satisfy the following SEM

$$\mathbf{y}_i = \mathbf{A}\mathbf{y}_i + \mathbf{B}\mathbf{x}_i + \mathbf{e}_i \quad (1)$$

where $\mathbf{A} = (a_{jj'}) \in \mathbb{R}^{p \times p}$ with zeros on the diagonal, $\mathbf{B} = (b_{jj'}) \in \mathbb{R}^{p \times q}$, $\mathbf{e}_i \sim N_p(0, \mathbf{\Sigma})$ with diagonal covariance matrix $\mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_p)$ and \mathbf{e}_i and \mathbf{x}_i are independent. Assuming that $\mathbf{I} - \mathbf{A}$ is invertible, model (1) can be written as

$$\mathbf{y}_i | \mathbf{x}_i \sim N_p \left\{ (\mathbf{I} - \mathbf{A})^{-1} \mathbf{B} \mathbf{x}_i, (\mathbf{I} - \mathbf{A})^{-1} \mathbf{\Sigma} (\mathbf{I} - \mathbf{A})^{-T} \right\}, \quad (2)$$

We define an RGM, $\mathcal{G} = (V, E)$, associated with the SEM (1) by letting vertices $V = \{1, \dots, p + q\}$ denote the variables in $(\mathbf{y}_i, \mathbf{x}_i)$. That is, the variables corresponding to \mathbf{y}_i are indexed $1, \dots, p$, and the variables corresponding to \mathbf{x}_i are indexed $p + 1, \dots, p + q$. We allow directed edges terminating in y -vertices, $j = 1, \dots, p$, and undirected edges between x -vertices. Specifically, we draw a directed edge $j' \rightarrow j$ from a y -vertex $j' = 1, \dots, p$, to another y -vertex $j = 1, \dots, p$ if $a_{jj'} \neq 0$. And we draw directed edges $h \rightarrow j$ from x -vertices $h = p + 1, \dots, p + q$ to y -vertices if $b_{j, h-p} \neq 0$. Finally, we draw undirected edges $h - h'$ between x -vertices, $h = p + 1, \dots, p + q$ and $h' = p + 1, \dots, p + q$, $h \neq h'$. The distribution of $(\mathbf{y}_i, \mathbf{x}_i)$ is global Markov with respect to the resulting RGM \mathcal{G} (Koster, 1996). The

construction defines a mapping between SEMs of the type (1) and RGMs. In this paper, \mathbf{B} is block diagonal because copy number and methylation of gene j , in principle, only affect the expression of gene j .

3 HRGMs with known groups

3.1 Probability model

Suppose the data arises from a heterogeneous population that is divided into K known groups. Let \mathbf{y}_k and \mathbf{x}_k be the $n_k \times p$ and $n_k \times q$ matrices of observations for group k , $k = 1, \dots, K$, with $n = \sum_{k=1}^K n_k$. We assume that the samples are exchangeable within each group and the probability model for subject i in group k is given by

$$\mathbf{y}_{ki} \mid \mathbf{x}_{ki}, \mathbf{A}^{(k)}, \mathbf{B}^{(k)}, \Sigma^{(k)} \sim \text{N}_p \left\{ (\mathbf{I} - \mathbf{A}^{(k)})^{-1} \mathbf{B}^{(k)} \mathbf{x}_{ki}, (\mathbf{I} - \mathbf{A}^{(k)})^{-1} \Sigma^{(k)} (\mathbf{I} - \mathbf{A}^{(k)})^{-T} \right\} \quad (3)$$

for $k = 1, \dots, K$ and $i = 1, \dots, n_k$ with $\mathbf{A}^{(k)} = (a_{jj'}^{(k)})$, $\mathbf{B}^{(k)} = (b_{jj'}^{(k)})$ and $\Sigma^{(k)} = \text{diag}(\sigma_1^{(k)}, \dots, \sigma_p^{(k)})$. Let $\boldsymbol{\theta}^{(k)} = (\mathbf{A}^{(k)}, \mathbf{B}^{(k)}, \Sigma^{(k)})$ denote the group-specific parameters. By the earlier described mapping between an SEM (1) and an RGM model, the $\boldsymbol{\theta}^{(k)}$ defines group-specific RGM models \mathcal{G}_k . The structural zeros of $\mathbf{A}^{(k)}$ and $\mathbf{B}^{(k)}$ correspond to missing edges in the RGM \mathcal{G}_k for group k .

Next we introduce a model feature to induce sparsity in $\mathbf{A}^{(k)}$ and $\mathbf{B}^{(k)}$. We use a non-local prior, defined as follows. First, we expand each entry as

$$a_{jj'}^{(k)} = \tilde{a}_{jj'}^{(k)} I(|\tilde{a}_{jj'}^{(k)}| > t_{jj'}) \quad \text{and} \quad b_{jj'}^{(k)} = \tilde{b}_{jj'}^{(k)} I(|\tilde{b}_{jj'}^{(k)}| > t_{jj'}) \quad (4)$$

where $\tilde{a}_{jj'}^{(k)}$ and $\tilde{b}_{jj'}^{(k)}$ are latent variables and the threshold parameters $t_{jj'}$ defines a minimum effect sizes of $a_{jj'}^{(k)}$ and $b_{jj'}^{(k)}$. That is, we represent $a_{jj'}^{(k)}$ and $b_{jj'}^{(k)}$ by thresholding latent $\tilde{a}_{jj'}^{(k)}$ and $\tilde{b}_{jj'}^{(k)}$. The advantages of this thresholding mechanism over standard Bayesian variable selection framework, for example, a spike-and-slab prior will become clear when we discuss the priors that link multiple groups. Notice that $t_{jj'}$ is edge-specific but shared across different groups, which is the key for inducing the desired dependence of graphs across similar groups and will be discussed in more detail in Section 3.2.

Models (3) and (4) involve three sets of parameters (i) effect sizes $\tilde{a}_{jj'}^{(k)}$ and $\tilde{b}_{jj'}^{(k)}$; (ii) thresholds t_j ; and (iii) variance matrices $\Sigma^{(k)}$.

3.2 Priors linking multiple groups

We first introduce the prior for $\tilde{a}_{jj'}^{(k)}$. We assume multivariate normal priors on $\tilde{\mathbf{a}}_{jj'} = \left(\tilde{a}_{jj'}^{(1)}, \dots, \tilde{a}_{jj'}^{(K)}\right)^T$,

$$\tilde{\mathbf{a}}_{jj'} \sim p(\tilde{\mathbf{a}}_{jj'}) = \text{N}(0, \tau_{jj'} \mathbf{\Omega}) \quad (5)$$

where $\tau_{jj'}$ is an edge-specific variance component. The $K \times K$ matrix $\mathbf{\Omega} = (\omega_{kk'})$ links edge strength (effect sizes) and by the thresholding in (4) also edge inclusion across K different groups, which in turn also links the graphs across groups. The magnitude of the off-diagonal entry $\omega_{kk'}$ of $\mathbf{\Omega}$ determines the strength of the correlation between groups k, k' . When $\omega_{kk'}$ is significantly away from zero, $\tilde{a}_{jj'}^{(k)}$ and $\tilde{a}_{jj'}^{(k')}$ are likely to be of similar magnitude *a priori* and since the thresholding parameters $t_{jj'}$ are shared across groups, there is a high probability that $\tilde{a}_{jj'}^{(k)}$ and $\tilde{a}_{jj'}^{(k')}$ are either both non-zero or both shrunk to zero. Therefore, graphs \mathcal{G}_k and $\mathcal{G}_{k'}$ are more likely to share common edges. On the other hand, when $\omega_{kk'}$ is negligible (i.e. close to 0), groups k and k' are unrelated. We do not constrain $\omega_{kk'}$ to be non-negative, which is imposed by Peterson et al. (2015). This additional flexibility allows edge strength to have different signs for different groups, which is desirable in estimating gene networks because potentially the gene regulations can switch from activation to inactivation across groups (e.g. case vs control). Note that $\tau_{jj'}$ and $\mathbf{\Omega}$ are not identifiable because $\tau_{jj'} \mathbf{\Omega} = c \tau_{jj'} \mathbf{\Omega} / c$ for any $c > 0$, which can be resolved by fixing $\Omega_{11} = 1$. The priors for $\tilde{b}_{jj'}^{(k)}$ are defined in a similar fashion,

$$\tilde{\mathbf{b}}_{jj'} = \left(\tilde{b}_{jj'}^{(1)}, \dots, \tilde{b}_{jj'}^{(K)}\right)^T \sim \text{N}(0, \lambda_{jj'} \mathbf{\Omega}). \quad (6)$$

3.3 Thresholding priors and induced marginals

We first discuss the condition of the prior for $t_{jj'}$ under which the induced marginal priors for $\mathbf{a}_{jj'} = \left(a_{jj'}^{(1)}, \dots, a_{jj'}^{(K)}\right)^T$ and $\mathbf{b}_{jj'} = \left(b_{jj'}^{(1)}, \dots, b_{jj'}^{(K)}\right)^T$ are mixtures of non-local priors

(Johnson and Rossell, 2010). For simplicity, we focus our discussion on $\mathbf{a}_{jj'}$ (the same argument holds for $\mathbf{b}_{jj'}$) and suppress the subscript jj' when it is clear from the context. Let $S \subseteq \{1, \dots, K\}$ denote the subset such that $a^{(k)} = 0$ if and only if $k \in S^c$. Let $\mathbf{a}_S = \{a^{(k)} \mid k \in S\}$ denote the group of non-zero coefficients. Then the prior for \mathbf{a} conditional on t can be written as a mixture over all possible S . Letting $C_S = \{|\tilde{a}^{(k)}| \leq t, k \in S^c\}$ denote the event that $|\tilde{a}^{(k)}| \leq t$ for $k \in S^c$, we have

$$p(\mathbf{a} \mid t) = \sum_{S \in 2^{\{1, \dots, K\}}} p(C_S \mid t) p_{\tilde{\mathbf{a}}_S}(\mathbf{a}_S \mid C_S, t) \prod_{k \in S} I(|a^{(k)}| > t) \prod_{k \in S^c} \delta_0(a^{(k)}),$$

where $2^{\{1, \dots, K\}}$ is the power set of $\{1, \dots, K\}$, $\tilde{\mathbf{a}}_S = \{\tilde{a}^{(k)} \mid k \in S\}$ and $p_{\tilde{\mathbf{a}}_S}(\cdot)$ is the distribution of the elements in $\tilde{\mathbf{a}}$ corresponding to non-zero elements of \mathbf{a} . We provide the marginal prior in the following lemma.

Lemma 1. *The marginal prior $p(\mathbf{a})$ is given by*

$$p(\mathbf{a}) = \sum_{S \in 2^{\{1, \dots, K\}}} w_{|S|} \times p_S(\mathbf{a}),$$

where $|S|$ is the cardinality of S , $w_{|S|}$ is a mixture weight, with $\sum_{S \in 2^{\{1, \dots, K\}}} w_{|S|} = 1$ and $p_S(\mathbf{a})$ (the actual form is provided in the Appendix) is the prior under the hypothesis $H_S : \mathbf{a}_S \neq \mathbf{0}$ and $\mathbf{a}_{S^c} = \mathbf{0}$. The weights $w_{|S|}$ in the mixture over S depend on S only indirectly through its cardinality. Moreover, $p_S(\mathbf{a})$ is a non-local alternative prior for any $S \in 2^{\{1, \dots, K\}} \setminus \emptyset$, that is, $p_S(\mathbf{a}) \rightarrow 0$ as $\mathbf{a}_S \rightarrow \mathbf{0}$, provided that $p(\tilde{\mathbf{a}})$ is bounded near $\mathbf{0}$ and $Pr(t = 0) = 0$.

The proof is given in the Appendix. Lemma 1 is a generalization of the result in Ni et al. (2016b) who show a similar result for a prior $p(\tilde{\mathbf{a}})$ with independent components. The nature of the marginal distribution as a non-local prior is the main motivation for introducing the construction in (4). Non-local priors (Johnson and Rossell, 2010) introduce multiple shrinkage, shrinking small effect to zero, which is useful in our setting as we are only interested in edges with significant strength. Since prior (5) is multivariate normal and hence bounded near $\mathbf{0}$, we only require $Pr(t = 0) = 0$ for $p_S(\mathbf{a})$ to be non-local; we assume a simple uniform prior for $t_{jj'} \sim \text{Unif}(0, b_t)$. We remark that the induced prior

for $\mathbf{A}^{(k)}$ should be restricted to the cone of invertible $\mathbf{I} - \mathbf{A}^{(k)}$, which practically is not a constraint given that any random matrix is almost surely invertible (Rudelson, 2008).

3.4 Priors for relatedness matrix and hyperparameters

We assign a conjugate inverse-Wishart prior for the relatedness matrix $\mathbf{\Omega}$ which controls the relatedness of different groups as discussed in Section 3.2,

$$\mathbf{\Omega} \sim \text{IW}(\nu, \mathbf{\Phi}),$$

with degrees of freedom ν and a scale matrix $\mathbf{\Phi}$. Alternatively, one could introduce a hyper MRF graph \mathcal{H} and put a G-Wishart prior on $\mathbf{\Omega}^{-1} \sim \text{W}_{\mathcal{H}}(\cdot, \cdot)$ if one wants to learn the conditional independence relationships between groups. Since in our case study, we only have $K = 3$ groups and the correlations between groups are quite significant (as shown in Section 6.1), we are not pursuing this direction in this paper. We complete the model by assuming conjugate priors $\sigma_j^{(k)} \sim \text{IG}(a_\sigma, b_\sigma)$, $\tau_{jj'} \sim \text{IG}(a_\tau, b_\tau)$ and $\lambda_{jj'} \sim \text{IG}(a_\lambda, b_\lambda)$. For our simulations and case studies, we use noninformative hyperparameters. But if desired, informative prior can be imposed for $\tau_{jj'}$ or $\lambda_{jj'}$ to encourage or discourage the inclusion of certain edges across all groups, which is useful, for example, when one wants to incorporate prior knowledge from a reference network.

4 HRGMs with unknown groups

We introduce a clustering approach to split heterogeneous samples into (unknown) homogeneous groups. Under a Bayesian approach, the number K of clusters need not be fixed *a priori* and can be inferred from the data. We first introduce a latent cluster membership indicator s_i with $s_i = k$ when subject i belongs to group k . Conditional on $s_i = k$ the likelihood is the same as (3). For reference we restate it here, now with conditioning on s_i ,

$$\mathbf{y}_i \mid \mathbf{x}_i, \boldsymbol{\theta}^{(k)}, s_i = k \sim \text{N}_p \left\{ (\mathbf{I} - \mathbf{A}^{(k)})^{-1} \mathbf{B}^{(k)} \mathbf{x}_i, (\mathbf{I} - \mathbf{A}^{(k)})^{-1} \mathbf{\Sigma}^{(k)} (\mathbf{I} - \mathbf{A}^{(k)})^{-T} \right\}. \quad (7)$$

We still need prior distributions for s_i and K . We use a Dirichlet-multinomial (DM) allocation model,

$$\begin{aligned} s_i | \boldsymbol{\pi}, K &\sim \text{Multinomial}(1, \pi_1, \dots, \pi_K), \\ \boldsymbol{\pi} | K &\sim \text{Dir}(\eta, \dots, \eta), \end{aligned}$$

and a geometric prior for $K \sim \text{Geo}(\rho)$. We refer to model (7) with a DM prior as the HRGM-DM.

Alternatively, we consider nonparametric Bayesian priors that give rise to exchangeable random partitions such as Poisson-Dirichlet process, also known as Pitman-Yor (PY) process priors (Pitman and Yor, 1997). A PY process induces a prior distribution on s_i 's and K which is characterized by a (modified) Chinese restaurant process $\text{CRP}(\alpha, d)$ with total mass parameter α and discount parameter d ,

$$p(s_i = k \mid s_1, \dots, s_{i-1}) \propto \begin{cases} n_k^{(-i)} - d & \text{for } k = 1, \dots, K^{(-i)} \\ \alpha + dK^{(-i)} & \text{for } k = K^{(-i)} + 1 \end{cases} \quad (8)$$

where $n_k^{(-i)}$ and $K^{(-i)}$ are the size of cluster k and the total number of clusters after removing the i th sample. The admissible values for (α, d) are $d \in [0, 1)$ with $\alpha > -d$ or $d < 0$ with $\alpha = |md|$ for some integer m . The popular Dirichlet process (Ferguson, 1973; Blackwell and MacQueen, 1973) is a special case of PY process when $d = 0$. The extra parameter d in PY process makes it more flexible than Dirichlet process prior (De Blasi et al., 2015). We refer to model (7) with the PY process prior as the HRGM-PY. De Blasi et al. (2015) and Barrios et al. (2013) argue for the PY prior as flexible prior for random partitions when it is desired to generalize the more restrictive assumptions of the DP prior and a parametric DM prior.

The choice of $\boldsymbol{\Omega}$ is motivated by the following consideration. The goal is to divide subjects into subpopulations with distinct networks, rather than induce dependence between clusters. We therefore set the relatedness matrix $\boldsymbol{\Omega} = \text{diag}(\Omega_1, \dots, \Omega_K)$ to be diagonal, i.e. networks are independent across clusters.

Implementing posterior inference under the HRGM-DM or the HRGM-PY model is

straightforward by MCMC simulation. Details are provided in Supplementary Material A.

The cluster membership indicators $\mathbf{s} = (s_1, \dots, s_n)$ describe a partition of $[n] = \{1, \dots, n\}$ into K clusters. A point estimate of the cluster arrangement can be evaluated with MCMC samples $\{s_i^{(\ell)} : i = 1, \dots, n; \ell = 1, \dots, L\}$ where the superscript (ℓ) denotes the ℓ -th posterior sample. We first define an $n \times n$ co-clustering matrix $\mathbf{C}^{(\ell)}$ for each posterior sample with entries $C_{ij}^{(\ell)} = I(s_i^{(\ell)} = s_j^{(\ell)})$. The posterior expectation $E(\mathbf{C} \mid \mathbf{y})$ is approximated by the ergodic element-wise average $\overline{\mathbf{C}} = \frac{1}{L} \sum_{\ell} \mathbf{C}^{(\ell)}$. Following the strategy in Dahl (2006), we define the least-squares estimate of the cluster arrangement as

$$\hat{\mathbf{C}} = \arg \min_{\mathbf{C} \in \{\mathbf{C}^{(1)}, \dots, \mathbf{C}^{(L)}\}} \|\mathbf{C} - \overline{\mathbf{C}}\|_F^2.$$

where $\|\cdot\|_F$ denotes Frobenius norm.

5 Simulation studies

In this section, we illustrate inference under the HRGM with known groups (Section 5.1) and unknown groups (Section 5.2) on synthetic data generated from RGMs with $p = 10$ and $q = 20$. In anticipation of the upcoming case study we refer to the first p variables as gene expression and the last q variables as DNA-level measurements, including two per gene. The hyperparameters are specified as $a_{\sigma} = b_{\sigma} = a_{\tau} = b_{\tau} = a_{\lambda} = b_{\lambda} = 0.01, b_t = 1, \nu = K, \Phi = \mathbf{I}$. The remaining parameters in the HRGM-DM and the HRGM-PY are set as $\eta = 1, \rho = 0.5, \alpha = 3, d = 0.5$. Each simulation is repeated 50 times to generate the following results.

5.1 Simulations with known groups

We consider two scenarios with different numbers of groups and group sizes: Under scenario 1 we use $K = 4, n_1 = n_2 = n_3 = n_4 = 50$; and under scenario 2, $K = 3, n_1 = 298, n_2 = 100, n_3 = 432$. Scenario 2 is designed to mimic the data in the upcoming application. We generate a simulation truth for $\tau_{jj'}$ as $\tau_{jj'} = 0.1$ with probability $1/5$ and $\tau_{jj'} = 10^{-6}$ with probability $4/5$. Similarly, we set $\lambda_{jj'} = 0.5$ with probability $3/4$ and $\lambda_{jj'} = 10^{-6}$ with probability $1/4$. Then $\tilde{\mathbf{a}}_{jj'}$ and $\tilde{\mathbf{b}}_{jj'}$ are generated from their respective prior distributions

(5) and (6) with $\mathbf{\Omega} = 0.5\mathbf{I}_K + 0.5\mathbf{J}_K$ where \mathbf{J}_K is a $K \times K$ matrix of ones, i.e. all correlations are set to 0.5. Then we threshold $\tilde{\mathbf{a}}_{jj'}$ and $\tilde{\mathbf{b}}_{jj'}$ at 0.2 to obtain $\mathbf{a}_{jj'}$ and $\mathbf{b}_{jj'}$. Each pair of graphs differ in about 10%–50% of edges. The hypothetical DNA-level measurements (that is, \mathbf{x}_i) are generated from independent standard normal distributions and gene expressions (\mathbf{y}_i) are generated from model (3) with $\mathbf{\Sigma}^{(k)} = 0.5\mathbf{I}$.

We compare inference under the HRGM model with a benchmark approach that estimates separate RGMs (Ni et al., 2016a) for each group. The following summaries are related to the classification of edges as included (“positive”) or not (“negative”). Table 1 reports the true positive rate (TPR), false discovery rate (FDR) and Matthews correlation coefficient (MCC). Inference under the HRGM compares favorably to independent analyses under separate RGMs for each group. The difference is most substantial in FDR. For example in scenario 1, FDR is well controlled at 0.07 for inference under the HRGM whereas under the separate RGMs the FDR is inflated to 0.40. Figures 1(a) and 1(b) show the receiving operating characteristic (ROC) curves, with the area under the ROC curve (AUC) listed in Table 1.

Table 1: Operating characteristics for inference under the HRGM and under separate RGMs for known groups. Standard deviations (across repeat simulations) are given in parentheses. Summaries refer to the estimation of edges as included versus excluded (MCC = misclassification rate, TPR = true positive rate, FDR = false discovery rate, AUC = area under the ROC curve).

Scenario	Method	MCC	TPR	FDR	AUC
1	HRGM	0.78(0.04)	0.75(0.04)	0.07(0.04)	0.93(0.02)
	RGM	0.48(0.04)	0.69(0.04)	0.40(0.05)	0.82(0.02)
2	HRGM	0.85(0.04)	0.90(0.03)	0.13(0.05)	0.97(0.01)
	RGM	0.75(0.06)	0.76(0.04)	0.15(0.09)	0.91(0.01)

5.2 Simulations with unknown groups

We again consider two scenarios similar to before, with scenario 2 mimicking the upcoming application. Under scenario 1, we use $K = 4, n_1 = n_2 = n_3 = n_4 = 50$. Under scenario 2 we use $K = 2, n_1 = 600, n_2 = 120$. The data are then generated in the same way as in

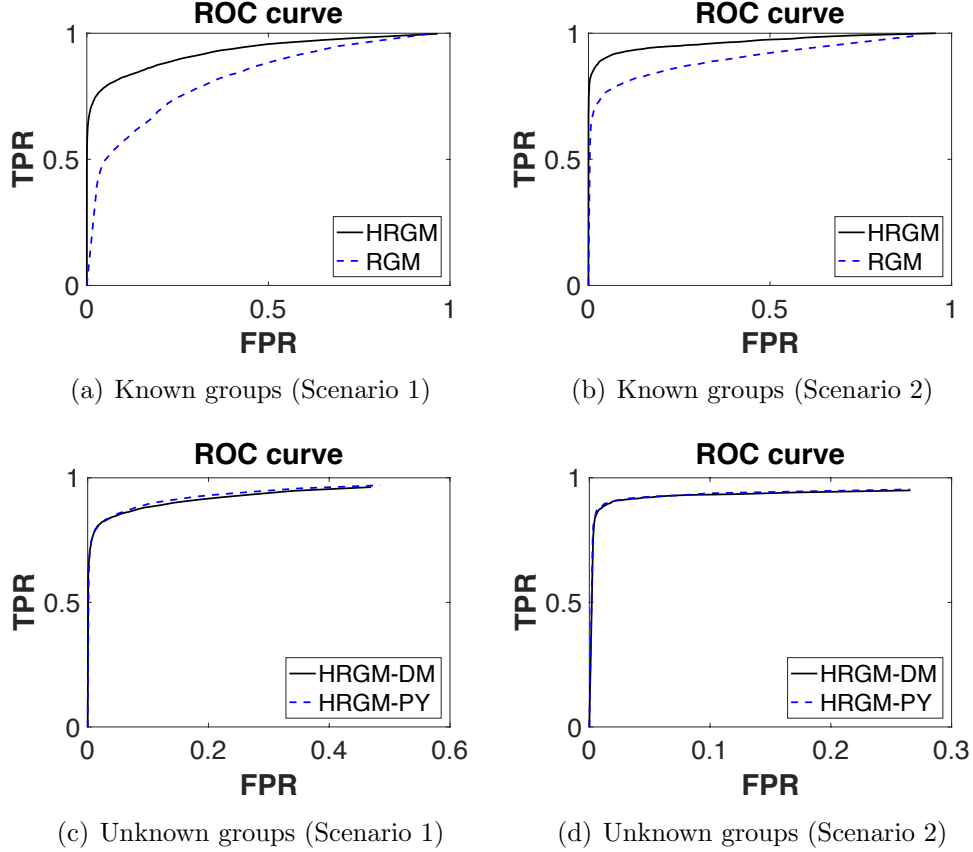


Figure 1: ROC curves. Top two figures: HRGM (solid) vs RGM (dashed) with known groups. Bottom two figures: HRGM-DM (solid) vs HRGM-PY (dashed) with unknown groups

Section 5.1, except that now we set $\Omega = \mathbf{I}$. We carry out inference under the HRGM-DM and the HRGM-PY conditional on the simulated data.

Table 2 shows summaries related to inference on the cluster arrangements. The table reports the percentage of simulations for which each method correctly identifies the true number of clusters. Let \hat{K} denote the number of estimated clusters that contain at least 5% of the samples. The table reports the probability (over repeated simulations) of $\hat{K} = K$. We also report the misclassification rate conditional on the true number K of clusters. In scenario 2, the two methods perform equally well. In scenario 1, however, while inference under the HRGM-DM always correctly estimates the true number of clusters, inference under the HRGM-PY fails in 30% of the simulations. This is probably due to the fact that the true cluster sizes are evenly distributed (50 samples each), but the PY prior favors a

small number of larger clusters and a large number of smaller clusters (Hjort et al., 2010). Compare also with the conditional cluster membership probabilities (8). For this reason, in the upcoming case studies, we only use the HRGM-DM model.

To evaluate the estimation of the networks, we again report TPR, FDR MCC and AUC as in Section 5.1. The summaries are evaluated conditional on the true number of clusters. We find acceptable operating characteristics for inference under both models, including the ROC curves in Figures 1(c) and 1(d).

Table 2: Operating characteristics for HRGM with unknown groups. Numerical standard deviations are in parentheses. The column labeled “ $\hat{K} = K$ ” reports the proportion of simulations which correctly estimate K . The remaining columns are as in Table 1. Rows labeled with “DM” report inference under the HRGM-DM and “PY” under the HRGM-PY.

Sc.	Model	$\hat{K} = K$	Misc	MCC	TPR	FDR	AUC
1	DM	1.00(0.00)	0.01(0.00)	0.83(0.03)	0.82(0.04)	0.08(0.04)	0.95(0.02)
	PY	0.70(0.46)	0.01(0.01)	0.83(0.05)	0.82(0.04)	0.08(0.06)	0.95(0.02)
2	DM	1.00(0.00)	0.01(0.00)	0.90(0.05)	0.89(0.04)	0.06(0.06)	0.96(0.02)
	PY	1.00(0.00)	0.01(0.00)	0.90(0.04)	0.89(0.03)	0.05(0.04)	0.96(0.02)

6 Case studies

6.1 p53 pathway across three cancer subtypes

A recent pan-cancer genomic study (Hoadley et al., 2014) has identified 11 major cancer subtypes based on molecular characterizations instead of tissue-of-origin. This provides independent information for clinical prognosis. Although the subtypes are correlated with tissue-of-origin, several distinct cancer types are classified into common subtypes. For example, head and neck squamous cell carcinoma (HNSC) and lung squamous cell carcinoma (LUSC) are classified into one subtype “C2-Squamous-like” by molecular alterations including p53 whereas breast invasive cancer (BRCA) is identified as subtype “C3-BRCA/Luminal” by itself.

Using inference under the proposed HRGM, we explore related dependencies of gene networks in the same three cancer types, HNSC, LUSC and BRCA. That is, we construct a

gene network for each cancer and borrow strength across cancer types adaptively, depending on how similar the cancers are. We expect the networks of HNSC and LUSC to have stronger association than those of HNSC and BRCA or of LUSC and BRCA since HNSC and LUSC belong to the same molecular subtype.

We use the software package TCGA-Assembler (Zhu et al., 2014) to retrieve mRNA gene expression, DNA copy number and DNA methylation for HNSC, LUSC and BRCA from The Cancer Genome Atlas (TCGA). We focus on genes that are mapped to the p53 pathway ($p = 10$) which responds to stresses that can disrupt the fidelity of DNA replication and cell division (Harris and Levine, 2005) and plays critical role in all the three cancers (Gasco et al., 2002; Leemans et al., 2011; Perez-Moreno et al., 2012). We take the samples of HNSC, LUSC and luminal BRCA analyzed in Hoadley et al. (2014) and match genomic data from different platforms, using methylation and copy number variation corresponding to each gene to record $q = 2p = 20$ DNA level variables. The resulting dataset includes $p + q = 30$ variables for $n_1 = 298$ HNSC samples, $n_2 = 100$ LUSC samples and $n_3 = 432$ luminal BRCA samples. We implement inference under the proposed HRGM model for known subgroups by posterior MCMC simulation. We run MCMC simulation for 100,000 iterations, discard the first 50% as burn-in and thin the chain to every 5th sample. MCMC diagnostics show no evidence for lack of practical convergence (see Supplementary Material B for details).

Figure 2 shows point estimates for the gene networks, based on controlling posterior expected FDR (Newton et al., 2004; Müller et al., 2006) at 1%. The edges between genes and corresponding copy number and methylation are omitted for clarity. These associations are shown separately in Table 3. Most gene expressions are associated with the corresponding copy number whereas only a few genes are correlated to their methylation. In Figure 2, edges that are shared across all three subtypes are represented as solid lines. Differential edges are represented as dashed lines. Arrowheads denote stimulatory regulations, whereas horizontal bars indicate inhibitory regulations. The number of edges in each of the three gene networks and the number of shared edges between each pair of networks are given

below,

$$\text{Number of shared edges} = \begin{pmatrix} & H & L & B \\ H & 16 & 10 & 9 \\ L & & 11 & 8 \\ B & & & 13 \end{pmatrix}.$$

Due to the molecular similarity of HNSC and LUSC, they share many edges: 10 out of 11 edges in the LUSC network are also found in the HNSC network. And as expected, BRCA shares fewer, but still a reasonable number of edges with HNSC and LUSC. In summary, inference under the proposed HRGM model recognizes difference in association between cancer types and borrows strength adaptively in a way that confirms the subtypes found by Hoadley et al. (2014). The estimated levels of adaptive borrowing of strength is reflected in the estimated relatedness matrix $\hat{\Omega}$,

$$\hat{\Omega} = \begin{pmatrix} & H & L & B \\ H & 1 & 0.77 & 0.63 \\ L & & 0.81 & 0.57 \\ B & & & 0.76 \end{pmatrix}.$$

The moderate to strong correlations suggest that the joint analysis under the HRGM is more appropriate than separate inference under separate RGMs.

A noticeable feature in all three estimated networks in Figure 2 is the central role of E2F1. E2F1 is an important transcription factor gene across cancer types which interacts with multiple genes in all networks. Such highly connected genes are known as hub genes and are often more involved in multiple regulatory activities than non-hub genes. In fact, E2F1 has a pivotal role in controlling cell cycle progression and induces apoptosis. It has been found that E2F1 is mutated in most, if not all, human tumors (Polager and Ginsberg, 2009). Our finding that E2F1 plays important roles in all networks is consistent with the fact that hub genes are more likely to be conserved across species, diseases and tumor (sub-)types (Casci, 2006). In addition, some edges that we find in our analysis are well studied in the biological literature. For example, ATM phosphorylates and stabilizes E2F1 in response to various stresses including DNA damage. Our study confirms this positive

regulatory relationship across all cancers.

Table 3: Association between mRNA gene expressions and DNA level measurements for HNSC, LUSC and BRCA. We use “c” and “m” to denote if the gene expression is associated with its copy number and methylation, respectively.

Cancer	Gene									
	TP53	ATM	CDKN1A	CDKN2A	CHEK1	CHEK2	E2F1	EP300	MDM2	MDM4
HNSC		c	c,m	c,m	c	c	c,m	c	c	c
LUSC			c	c,m	c	c	c,m	c	c	c
BRCA	c	c		c	c	c	c	c	c	c

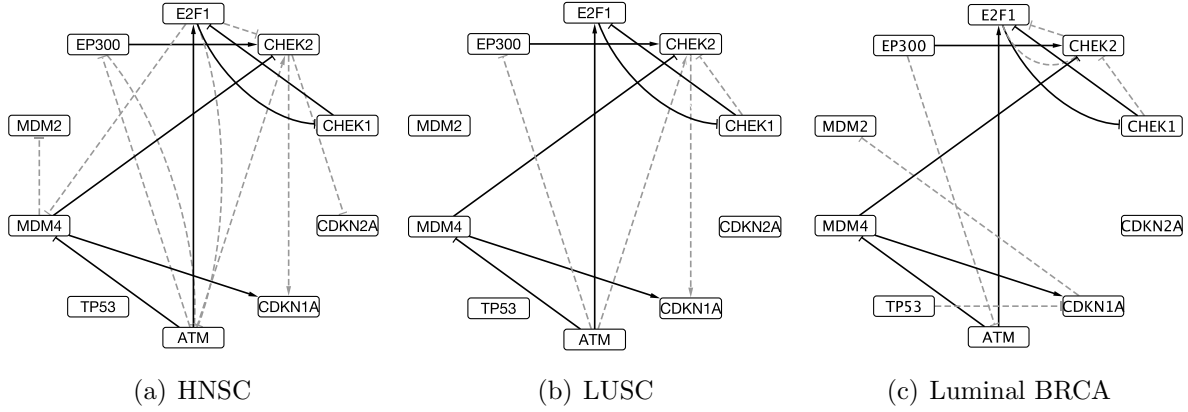


Figure 2: Inferred gene networks for HNSC, LUSC and luminal BRCA. Shared edges across all subtypes are represented by solid lines; differential edges are shown as dashed lines. Arrowheads represent stimulatory interactions, whereas horizontal bars denote inhibitory regulations.

6.2 Clustering breast cancer subtypes

Breast cancer is known to be a highly heterogeneous disease. Based on three biomarkers (ER/PR/HER2), breast cancer is traditionally classified into three groups: luminal, HER2 and basal. However, this classification system is also known to be suboptimal and in need of improvement for better diagnostics and prognostics (Di Leo et al., 2015).

Instead of three biomarkers, we consider a critical pathway, RAS-MAPK ($p = 10$ genes), which transmits and amplifies signals involved in cell proliferation and cell death in breast cancer (Santen et al., 2002). We use inference under the proposed HRGM-DM to find

subtypes with respect to the network of these 10 genes, based on breast cancer data ($n = 720$ samples) retrieved from TCGA as described in Section 6.1. Our goal is to simultaneously cluster patients and estimate gene networks for each cluster. We implement inference using MCMC posterior simulation, using the same setup as in Section 6.1. MCMC diagnostics show no evidence for lack of practical convergence (see Supplementary Material B for details).

We identify two major clusters with 600 and 107 samples. To explore the clinical relevance of the reported clusters we evaluate Kaplan-Meier (KM) estimates of patients' survivals for the two clusters. These are shown in Figure 3(a). We find a substantial difference in median survival between the two clusters: a difference of 928 days. For comparison, we also implemented inference under RLD (Rodriguez et al., 2011) and with the K-means algorithm with $K = 2$ for the same data. Their KM estimators are shown in Figures 3(b) and 3(c). RLD finds one major cluster with 687 patients. The second largest cluster has only 22 patients. The difference of median survival is 120.5 days between the two clusters. The two clusters identified by K-means have 635 and 85 samples with median survival differs by 399 days. In addition, we also compute KM estimates, shown in Figure 3(d), for the luminal/HER2 group versus basal group; the latter usually has much worse prognosis. The difference of median survival is 497 days.

To formally compare the clinical relevance in terms of significantly different overall survival for the reported clusters, we carried out log-rank tests (p-values shown in Figure 3) for the difference of the corresponding survival distributions. In fact, only HRGM-DM identifies clusters that have significantly different survival distributions. The HRGM-DM model detects the most distinctive clusters in terms of prognosis. Such clusters or subtypes may be useful in refining subtypes of breast cancer and developing new treatment options.

In terms of the estimated network structure, we find that all genes are associated with their respective copy number for both clusters except for *SOS1* and *KRAS* in cluster 1 and *SOS1* and *MAPK3* in cluster 2. In addition, *NRAS* and *KRAS* in both clusters and *SOS2* and *BRAF* in cluster 2 are found to be associated with their methylation. Figure 4 shows the estimated gene networks for the two clusters when controlling posterior expected FDR at 1%. As before, shared edges across both clusters are represented by solid lines and differ-

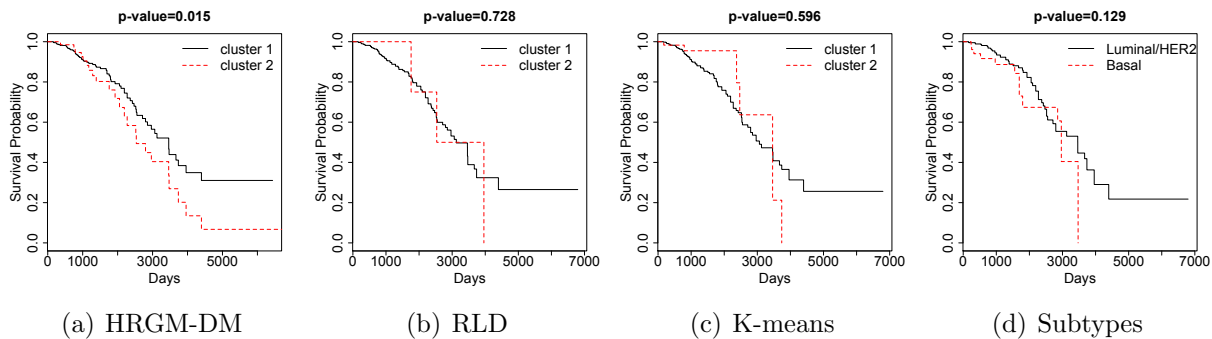


Figure 3: Kaplan-Meier estimators for cluster 1 (solid) and cluster 2 (dashed). P-values of log-rank test are shown on the top of each figure.

ential edges are shown as dashed lines. Arrowheads denote stimulatory regulations, whereas horizontal bars indicate inhibitory regulations. We find 27 edges for cluster 1 and 41 edges for cluster 2. The two networks share 18 edges but are otherwise quite different from each other. For example, the well-known cascade $SOS1 \rightarrow KRAS \rightarrow MAPK1 \rightarrow MAP2K2$ (Santen et al., 2002) is found in cluster 2 but not in cluster 1. Furthermore, we find $MAP2K2$ inhibits $SOS1$ in cluster 2, which completes the negative feedback loop, a commonly observed motif in gene network (Krishna et al., 2006), as shown in Figure 4(c). Interestingly, it is recently discovered that $MAP2K2$ phosphorylates and inhibits SOS , therefore reducing $MAP2K2$ activation (Mendoza et al., 2011).

7 Discussion

We have developed the HRGM(-DM,-PY) model for inference on gene networks for heterogeneous samples. The HRGM uses a multivariate normal prior to connect graphs across known groups by introducing correlation on edge strength and edge inclusion. In the case of unknown groups, the HRGM together with a Dirichlet-multinomial or Pitman-Yor process prior for the cluster arrangement allows to learn the unknown groups and estimate group-specific networks at the same time. Bayesian model selection is implemented with a thresholding prior and applied to obtain sparse network. Simulation studies demonstrate the advantages of HRGM over a comparable approach with separate inference for group-specific networks. In the first application, we find both common and differential network

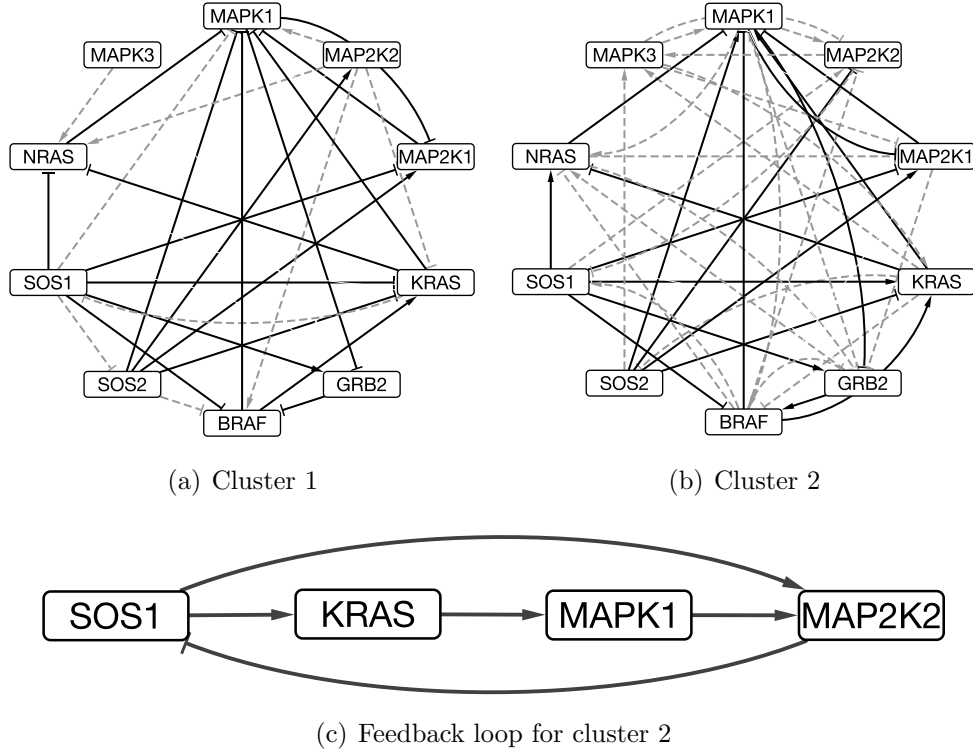


Figure 4: Inferred gene networks for breast cancer. Top two figures: networks for clusters 1 and 2 with shared edges represented by solid lines and differential edges represented by dashed lines. Bottom figure: negative feedback loop of gene network for cluster 2. Arrowheads represent stimulatory interactions, whereas horizontal bars denote inhibitory regulations.

structures for different cancer types. In the second application, we are able to identify clusters that differ significantly in network structures. The clinical significance of the discovered clusters is validated by significantly different cluster-specific survival functions. Inference may be informative in refining subtypes of breast cancer and developing new therapeutic options.

References

- Barrios, E., Nieto-Barajas, L. E., and Prünster, I. (2013). A study of normalized random measures mixture models. *Statistical Science*, page to appear.
- Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes.

- The annals of statistics*, pages 353–355.
- Cai, X., Bazerque, J. A., and Giannakis, G. B. (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Comput Biol*, 9(5):e1003068.
- Caspi, T. (2006). Network fundamentals, via hub genes. *Nature Reviews Genetics*, 7(9):664–665.
- Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, pages 201–218.
- Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397.
- De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE transactions on pattern analysis and machine intelligence*, 37(2):212–229.
- Di Leo, A., Curigliano, G., Diéras, V., Malorni, L., Sotiriou, C., Swanton, C., Thompson, A., Tutt, A., and Piccart, M. (2015). New approaches for improving outcomes in breast cancer in Europe. *The Breast*, 24(4):321–330.
- Dobra, A., Lenkoski, A., and Rodriguez, A. (2012). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.
- Gasco, M., Shami, S., and Crook, T. (2002). The p53 pathway in breast cancer. *Breast Cancer Research*, 4(2):70.
- Green, P. J. and Thomas, A. (2013). Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika*, 100(1):91–110.

- Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, page asq060.
- Harris, S. L. and Levine, A. J. (2005). The p53 pathway: positive and negative feedback loops. *Oncogene*, 24(17):2899–2908.
- Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian nonparametrics*, volume 28. Cambridge University Press.
- Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D., Niu, B., McLellan, M. D., Uzunangelov, V., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944.
- Ickstadt, K., Bornkamp, B., Grzegorzczak, M., Wieczorek, J., Sheriff, M. R., Grecco, H. E., and Zamir, E. (2011). Nonparametric bayesian networks. *Bayesian Statistics 9*, 9:283.
- Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170.
- Koster, J. T. (1996). Markov properties of nonrecursive causal models. *The Annals of Statistics*, 24(5):2148–2177.
- Krishna, S., Andersson, A. M., Semsey, S., and Sneppen, K. (2006). Structure and function of negative feedback loops at the interface of genetic and metabolic networks. *Nucleic acids research*, 34(8):2455–2462.
- Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558.
- Leemans, C. R., Braakhuis, B. J., and Brakenhoff, R. H. (2011). The molecular biology of head and neck cancer. *Nature Reviews Cancer*, 11(1):9–22.
- Lijoi, A., Nipoti, B., and Prünster, I. (2014). Dependent mixture models: clustering and borrowing information. *Computational Statistics & Data Analysis*, 71:417–433.

- Logsdon, B. A. and Mezey, J. (2010). Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Comput Biol*, 6(12):e1001014.
- Mendoza, M. C., Er, E. E., and Blenis, J. (2011). The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation. *Trends in biochemical sciences*, 36(6):320–328.
- Mitra, R., Müller, P., Ji, Y., et al. (2016). Bayesian graphical models for differential pathways. *Bayesian Analysis*, 11(1):99–124.
- Mukherjee, C. and Rodriguez, A. (2015). GPU-powered Shotgun Stochastic Search for Dirichlet process mixtures of Gaussian Graphical Models. *Journal of Computational and Graphical Statistics*, (just-accepted):00–00.
- Müller, P., Parmigiani, G., and Rice, K. (2006). FDR and Bayesian multiple comparisons rules.
- Müller, P., Quintana, F., and Rosner, G. L. (2012). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*.
- Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176.
- Ni, Y., Ji, Y., and Mueller, P. (2016a). Reciprocal graphical models for integrative gene regulatory network analysis. *arXiv preprint arXiv:1607.06849*.
- Ni, Y., Stingo, F., and Baladandayuthapani, V. (2016b). Bayesian graphical regression. *submitted*.
- Oates, C. J., Smith, J. Q., and Mukherjee, S. (2016). Estimating causal structure using conditional DAG models. *Journal of Machine Learning Research*, 17(54):1–23.
- Oates, C. J., Smith, J. Q., Mukherjee, S., and Cussens, J. (2015). Exact estimation of multiple directed acyclic graphs. *Statistics and Computing*, pages 1–15.

- Perez-Moreno, P., Brambilla, E., Thomas, R., and Soria, J.-C. (2012). Squamous cell carcinoma of the lung: molecular subtypes and therapeutic opportunities. *Clinical Cancer Research*, 18(9):2443–2451.
- Peterson, C., Stingo, F. C., and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174.
- Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900.
- Polager, S. and Ginsberg, D. (2009). p53 and E2f: partners in life and death. *Nature Reviews Cancer*, 9(10):738–748.
- Quintana, F. A. (2006). A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference*, 136(8):2407–2429.
- Rodriguez, A., Lenkoski, A., and Dobra, A. (2011). Sparse covariance estimation in heterogeneous samples. *Electronic journal of statistics*, 5:981.
- Rudelson, M. (2008). Invertibility of random matrices: norm of the inverse. *Annals of Mathematics*, pages 575–600.
- Santen, R. J., Song, R. X., McPherson, R., Kumar, R., Adam, L., Jeng, M.-H., and Yue, W. (2002). The role of mitogen-activated protein (MAP) kinase in breast cancer. *The Journal of steroid biochemistry and molecular biology*, 80(2):239–256.
- Wang, H. and West, M. (2009). Bayesian analysis of matrix normal graphical models. *Biometrika*, 96(4):821–834.
- Wang, W., Baladandayuthapani, V., Holmes, C. C., and Do, K.-A. (2013). Integrative network-based Bayesian analysis of diverse genomics data. *BMC Bioinformatics*, 14(Suppl 13):S8.
- Yajima, M., Telesca, D., Ji, Y., and Müller, P. (2015). Detecting differential patterns of interaction in molecular pathways. *Biostatistics*, 16(2):240–251.

Zhang, L. and Kim, S. (2014). Learning gene networks under SNP perturbations using eQTL datasets. *PLoS Comput Biol*, 10(2):e1003420.

Zhu, Y., Qiu, P., and Ji, Y. (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nature methods*, 11(6):599–600.

A

A.1 Proof of Lemma 1

We integrate t out from joint distribution $p(\mathbf{a}, t) = p(\mathbf{a}|t)p(t)$,

$$\begin{aligned}
p(\mathbf{a}) &= \int p(\mathbf{a}|t)p(t)dt \\
&= \sum_{S \in 2^{\{1, \dots, K\}}} \int_0^{\min_{k \in S} |a_k|} Pr(|\tilde{a}_k| \leq t, k \in S^c | t) p_{\tilde{a}_S}(a_S | |\tilde{a}_k| \leq t, k \in S^c, t) p(t) dt \prod_{k \in S^c} \delta_0(a_k) \\
&= \sum_{S \in 2^{\{1, \dots, K\}}} E_t[Pr(|\tilde{a}_k| \leq t, k \in S^c, |\tilde{a}_k| > t, k \in S | t)] \\
&\quad \times \frac{\int_0^{\min_{k \in S} |a_k|} Pr(|\tilde{a}_k| \leq t, k \in S^c | t) p_{\tilde{a}_S}(a_S | |\tilde{a}_k| \leq t, k \in S^c, t) p(t) dt \prod_{k \in S^c} \delta_0(a_k)}{E_t[Pr(|\tilde{a}_k| \leq t, k \in S^c, |\tilde{a}_k| > t, k \in S | t)]} \\
&:= \sum_{S \in 2^{\{1, \dots, K\}}} w_{|S|} \times p_S(\mathbf{a}).
\end{aligned}$$

If $p(\tilde{\mathbf{a}})$ is bounded near $\mathbf{0}$ and $Pr(t = 0) = 0$, the integrand on the numerator is bounded near $t = 0$. Since the upper limit of the integral goes to zero as $\mathbf{a}_S \rightarrow 0$, we obtain $p_S(\mathbf{a}) \rightarrow 0$ as $\mathbf{a}_S \rightarrow 0$.